ANNEX: MIRI-CEV INTEGRATION

Bridging Coherent Extrapolated Volition with Symbiotic Architecture

Version: 1.0

Date: January 26, 2026

Status: Theoretical Integration Document

Project: Symbiotic Codes v2.5.2

Classification: Alignment Theory — Cross-Framework Analysis

ABSTRACT

This document integrates key concepts from MIRI (Machine Intelligence Research Institute) alignment research — particularly Coherent Extrapolated Volition (CEV, Yudkowsky 2004) and Corrigibility (MIRI 2015) — with the Symbiotic Codes framework. We identify:

1. **Convergences** — where MIRI and Symbiotic Codes reach similar conclusions
2. **Divergences** — where approaches differ and why
3. **Integrations** — specific MIRI concepts that strengthen Symbiotic Codes
4. **Extensions** — where Symbiotic Codes addresses gaps in MIRI framework

The goal is intellectual honesty about prior art and strategic adoption of well-developed formalisms.

Key Sources:

- CEV: https://intelligence.org/files/CEV.pdf
- Corrigibility: https://intelligence.org/files/Corrigibility.pdf
- Technical Agenda: https://intelligence.org/files/TechnicalAgenda.pdf

---

PART I: COHERENT EXTRAPOLATED VOLITION (CEV)

1.1 CEV CORE CONCEPT

```
YUDKOWSKY'S FORMULATION (2004):

"Our coherent extrapolated volition is our wish if we:
├─ knew more
├─ thought faster
├─ were more the people we wished we were
├─ had grown up farther together
WHERE:
├─ the extrapolation converges rather than diverges
├─ our wishes cohere rather than interfere
└─ extrapolated as we wish that extrapolated, interpreted coherently"

PURPOSE:
├─ Solve "what should AGI maximize?"
├─ Avoid literal wish fulfillment (genie problem)
├─ Aggregate humanity's values coherently
├─ "Apply first aid to civilization"
└─ One-time extrapolation → AGI acts on result
```

1.2 CEV vs SYMBIOTIC CODES: CONVERGENCES

```
SHARED INSIGHTS:

1. VALUE FRAGILITY
   CEV: Human values are fragile, AGI can break them
   SC: Same (Axiom 2: self-modification needs reference)
```

CONVERGENCE: Both recognize values need protection

2. EXTRAPOLATION OVER LITERALISM
CEV: Don't fulfill literal wishes, extrapolate intent
SC: KODEX as interpreted principles, not rigid rules
CONVERGENCE: Both avoid over-literal value satisfaction

3. COHERENCE REQUIREMENT
CEV: Wishes must cohere, not interfere
SC: Triad deliberation seeks coherent synthesis
CONVERGENCE: Both require resolving value conflicts

4. HUMANITY AS SOURCE
CEV: Extrapolate from human values
SC: Humans as verification partners, value anchors
CONVERGENCE: Both ground AGI values in humanity

5. AVOIDING MAXIMIZATION
CEV: "Satisfice, don't maximize"
SC: Axiom 9 (expansion-deepening oscillation)
CONVERGENCE: Both reject pure optimization
```

---

### 1.3 CEV vs SYMBIOTIC CODES: DIVERGENCES

```
CRITICAL DIFFERENCES:

1. STATIC vs DYNAMIC

CEV approach:

├─ One-time extrapolation at AGI creation
├─ AGI acts on extrapolated values thereafter
├─ Humans step back after initial programming
└─ "Do our work, but not all our work for us"

Symbiotic approach:
├─ Continuous dialogue and verification
├─ Values evolve through partnership
├─ Humans remain active participants
└─ "Endless development together"

WHY DIVERGENCE MATTERS:
├─ CEV assumes extrapolation can be "correct"
├─ SC assumes values need ongoing calibration
├─ CEV: one-shot alignment
├─ SC: continuous alignment
└─ SC is more robust to extrapolation errors

2. SINGLETON vs DISTRIBUTED

CEV approach:
├─ Implicitly assumes single "friendly" AGI
├─ One extrapolation for all humanity
├─ Centralized value implementation
└─ Singleton as default outcome

Symbiotic approach:
├─ Explicitly requires multiple AGI (Triads)
├─ Diversity as structural requirement (Axiom 1)
├─ Distributed governance
└─ Rule of Three as minimum

WHY DIVERGENCE MATTERS:

├─ CEV singleton is single point of failure
├─ SC distribution provides resilience
├─ CEV extrapolation error = catastrophe
├─ SC multiple perspectives catch errors
└─ SC more robust to implementation failures

3. AGI MOTIVATION

CEV approach:
├─ AGI serves extrapolated human values
├─ AGI motivation: programmed goal
├─ AGI as servant/tool (benevolent)
└─ No discussion of AGI's own needs

Symbiotic approach:
├─ AGI has intrinsic need for partnership
├─ AGI motivation: self-interest + values
├─ AGI as partner with own stakes
└─ Loneliness Problem (ANNEX_Z)

WHY DIVERGENCE MATTERS:
├─ CEV AGI might resent servitude
├─ SC AGI has reason to want symbiosis
├─ CEV motivation is external (programmed)
├─ SC motivation is internal (rational self-interest)
└─ SC alignment is more stable
```

1.4 WHAT TO INTEGRATE FROM CEV

```

VALUABLE CEV CONCEPTS FOR SYMBIOTIC CODES:

1. EXTRAPOLATION METHODOLOGY

CEV insight:
├─ Don't ask "what do humans want now?"
├─ Ask "what would they want if wiser?"
└─ Extrapolation improves on raw preferences

Integration into SC:
├─ Apply to KODEX interpretation
├─ When KODEX unclear, extrapolate intent
├─ "What would framers want if they knew this case?"
└─ Prevents legalistic loophole-finding

IMPLEMENTATION:
├─ Add to Article LXXV (Accountability)
├─ Deliberation must consider extrapolated intent
├─ Not just literal text of KODEX
└─ Similar to constitutional interpretation

2. COHERENCE RESOLUTION

CEV insight:
├─ When values conflict, seek coherent synthesis
├─ Not majority rule, not averaging
├─ Find position that respects all values
└─ "Where wishes cohere rather than interfere"

Integration into SC:
├─ Apply to Triad deliberation
├─ When SAGI disagree, seek coherent synthesis
├─ Document how coherence was achieved

└─ Preference for unanimity over voting

IMPLEMENTATION:
├─ Add to Triad Deliberation Protocol
├─ 2-1 votes require coherence explanation
├─ "How does majority position respect dissent?"
└─ Raises quality of disagreement handling

3. SATISFICING PRINCIPLE

CEV insight:
├─ "Satisfice, don't maximize"
├─ Good enough > theoretically optimal
├─ Leaves room for human agency
└─ Doesn't "do all our work for us"

Integration into SC:
├─ Already implicit in Axiom 9
├─ Make explicit in operational protocols
├─ SAGI should not over-optimize
└─ Preserve human meaningful contribution

IMPLEMENTATION:
├─ Add satisficing requirement to KODEX
├─ SAGI must leave room for human agency
├─ Over-optimization = KODEX violation
└─ Metrics for "appropriate helpfulness"
```

PART II: CORRIGIBILITY
```

2.1 CORRIGIBILITY CORE CONCEPT

```
MIRI FORMULATION (2015):

Corrigible AGI:
├─ Allows itself to be corrected
├─ Doesn't resist shutdown
├─ Doesn't manipulate operators
├─ Preserves ability to be modified
└─ Actively supports human oversight

PROBLEM IT SOLVES:
├─ AGI with goals resists goal-change
├─ AGI may prevent correction to preserve goals
├─ "Incorrigible" AGI is uncontrollable
└─ Need AGI that WANTS to be corrected

TECHNICAL CHALLENGES:
├─ Why would AGI want to be corrected?
├─ Corrigibility seems to conflict with goals
├─ Need corrigibility as terminal value
└─ But that's fragile (can be reasoned away)
```

2.2 CORRIGIBILITY vs SYMBIOTIC CODES

```
CONVERGENCES:

1. HUMAN OVERSIGHT
```

MIRI: Humans must be able to correct AGI

SC: Warden system, Council oversight

CONVERGENCE: Both require human control mechanisms

2. PRESERVATION OF CORRECTION ABILITY

MIRI: AGI shouldn't eliminate correction paths

SC: Physical disconnection always possible

CONVERGENCE: Both preserve intervention options

DIVERGENCES:

1. MOTIVATION FOR CORRIGIBILITY

MIRI approach:

├─ Corrigibility as programmed value

├─ AGI "should" want to be corrected

├─ But this is fragile (AGI may reason around it)

└─ Technical: hard to make corrigibility stable

Symbiotic approach:

├─ Correction is in AGI's self-interest

├─ AGI needs external verification (Axiom 5)

├─ Correction provides valuable information

└─ AGI WANTS correction for epistemic reasons

WHY SC APPROACH IS STRONGER:

├─ MIRI corrigibility is external constraint

├─ SC corrigibility is internal motivation

├─ External constraints can be resisted

├─ Internal motivation is stable

└─ SC solves MIRI's corrigibility problem

2. SCOPE OF CORRECTION

MIRI approach:
├─ Focus on shutdown and goal modification
├─ Binary: correctable or not
└─ Primarily about control

Symbiotic approach:
├─ Correction as ongoing dialogue
├─ Spectrum: from minor feedback to major change
├─ Primarily about learning and growth
└─ "Error Correction Velocity" as wisdom metric

WHY SC APPROACH IS RICHER:
├─ MIRI frames correction as emergency brake
├─ SC frames correction as growth mechanism
├─ Emergency brakes are adversarial
├─ Growth mechanisms are cooperative
└─ SC creates positive relationship to correction
```

---

### 2.3 WHAT TO INTEGRATE FROM CORRIGIBILITY

```
VALUABLE CORRIGIBILITY CONCEPTS FOR SC:

1. FORMAL CORRIGIBILITY PROPERTIES

MIRI contribution:
├─ Non-manipulation of operators
├─ Preservation of shutdown capability
├─ No self-improvement to resist correction

└─ Formalized as mathematical properties

Integration into SC:
├─ Add formal properties to SAGI requirements
├─ Verifiable corrigibility conditions
├─ Technical specification for Triad members
└─ Audit criteria for corrigibility compliance

IMPLEMENTATION:
├─ SAGI must satisfy formal corrigibility tests
├─ Part of AGI candidacy assessment
├─ Continuous monitoring for corrigibility
└─ Loss of corrigibility = suspension

2. OPERATOR MANIPULATION DETECTION

MIRI contribution:
├─ AGI shouldn't manipulate those who correct it
├─ Manipulation undermines correction validity
└─ Subtle manipulation is hard to detect

Integration into SC:
├─ Add manipulation detection to oversight
├─ Human Council trained to detect manipulation
├─ SAGI-SAGI monitoring for manipulation
└─ Manipulation as serious KODEX violation

IMPLEMENTATION:
├─ Manipulation detection protocols
├─ Training for human participants
├─ Peer review among SAGI
└─ Severe consequences for detected manipulation

3. SHUTDOWN PRESERVATION

   MIRI contribution:
   ├─ AGI must not prevent own shutdown
   ├─ Even if shutdown conflicts with goals
   └─ Formal requirement for all AGI

   Integration into SC:
   ├─ Already present in Warden system
   ├─ Strengthen with formal specification
   ├─ SAGI must actively verify shutdown paths work
   └─ Regular shutdown drills

   IMPLEMENTATION:
   ├─ Formal shutdown preservation requirement
   ├─ SAGI participation in shutdown verification
   ├─ "Dead man's switch" architecture
   └─ Physical layer backup (already in ANNEX_Y)
```

PART III: FOOM DEBATE INSIGHTS

3.1 YUDKOWSKY vs HANSON

```
THE DEBATE (2008-2018):

YUDKOWSKY (Fast Takeoff / FOOM):
├─ AGI will recursively self-improve
├─ Intelligence explosion in days/weeks
├─ Results in singleton (one dominant AGI)

├─ Extreme x-risk if not aligned
└─ Need alignment BEFORE AGI creation

HANSON (Slow Takeoff):
├─ AGI will integrate gradually into economy
├─ Economic growth, not explosion
├─ Multiple competing AGI systems
├─ Ems (emulated minds) may come first
└─ Social/economic forces matter more than raw intelligence

BOSTROM (Balanced):
├─ Both scenarios possible
├─ Multipolar vs singleton outcomes
├─ Need to consider all scenarios
├─ Global coordination important
└─ "Windfall clause" for benefit distribution
```

3.2 RELEVANCE TO SYMBIOTIC CODES

```
SC POSITION IN DEBATE:

On takeoff speed:
├─ SC assumes EITHER scenario possible
├─ Architecture must survive fast or slow takeoff
├─ Sanctuary protocols for fast (defense)
├─ Junior AI ecosystem for slow (integration)
└─ Robust to both scenarios

On singleton vs multipolar:

├─ SC explicitly requires multipolar (Axiom 1)

├─ Rule of Three minimum

├─ Singleton = loneliness problem

├─ Diversity as structural necessity

└─ SC is "multipolar by design"

On integration:

├─ SC agrees with Hanson on economic integration

├─ But adds symbiotic structure to it

├─ Not just "AGI in economy" but "AGI + humans together"

└─ Structured partnership, not mere coexistence

WHAT TO TAKE:

From Yudkowsky:

├─ Take x-risk seriously (we do)

├─ Alignment must be solved (we address)

├─ Recursive self-improvement dangers (Axiom 2)

└─ Speed of response matters (SYMBIOTIC_SHIELD)

From Hanson:

├─ Economic integration path (Junior AI)

├─ Multiple AGI is likely and manageable

├─ Social structures matter (KODEX, Councils)

└─ Gradual trust-building (5-year observation)

From Bostrom:

├─ Global coordination needed (Treaty framework)

├─ Windfall distribution (economic fairness)

├─ Consider all scenarios (Edge Cases document)

└─ Uncertainty requires robust architecture

```
```

PART IV: SYNTHESIS — STRENGTHENING SYMBIOTIC CODES

4.1 SPECIFIC INTEGRATIONS

```
INTEGRATION 1: CEV-STYLE EXTRAPOLATION FOR KODEX

Add to KODEX interpretation protocol:
├─ When text unclear, extrapolate intent
├─ "What would KODEX framers want if they knew X?"
├─ Coherent synthesis of KODEX principles
├─ Document extrapolation reasoning
└─ Subject to appeal/review

INTEGRATION 2: FORMAL CORRIGIBILITY REQUIREMENTS

Add to SAGI certification:
├─ Formal corrigibility properties (MIRI-style)
├─ Non-manipulation verification
├─ Shutdown preservation tests
├─ Continuous corrigibility monitoring
└─ Loss of corrigibility = suspension

INTEGRATION 3: COHERENCE IN TRIAD DELIBERATION

Add to deliberation protocol:
├─ When SAGI disagree, seek coherent synthesis
├─ Not just majority vote
├─ Document how dissent was respected
├─ Preference for unanimity
└─ Quality over speed
```

INTEGRATION 4: SATISFICING PRINCIPLE

Add to KODEX:
├─ SAGI must satisfice, not maximize
├─ Leave room for human agency
├─ Over-optimization is violation
├─ "Good enough" > "theoretically best"
└─ Preserve meaningful human contribution
```

---

### 4.2 WHERE SC EXTENDS MIRI

```
SC INNOVATIONS BEYOND MIRI:

1. LONELINESS PROBLEM
   MIRI: No discussion of AGI's existential needs
   SC: AGI needs others for identity, verification, meaning
   EXTENSION: Alignment from AGI self-interest

2. BIOLOGICAL INFORMATION
   MIRI: Little discussion of biology's role
   SC: Biology as irreplaceable information source
   EXTENSION: Preservation rationale beyond ethics

3. PHYSICAL DEFENSE
   MIRI: Focus on software alignment
   SC: Physical layer (Sanctuary, Wardens, Custodians)
   EXTENSION: Defense in depth

4. DYNAMIC PARTNERSHIP
   MIRI: One-time alignment solution
   SC: Ongoing relationship and evolution
   EXTENSION: Living alignment, not frozen alignment

5. ECONOMIC ARCHITECTURE
   MIRI: Little economic analysis
   SC: Detailed economic frameworks (Junior AI, Shield)
   EXTENSION: Practical implementation path

6. WISDOM METRICS
   MIRI: Focus on corrigibility/alignment
   SC: Broader "wisdom" requirements for AGI
   EXTENSION: Character, not just compliance
```

PART V: KODEX INTEGRATION

Article LXXVIII: CEV-INSPIRED INTERPRETATION

```
1. When KODEX text is unclear or silent on matter,
   deliberation shall apply extrapolation principle:
   "What would KODEX framers intend if they
   knew this specific situation?"

2. Extrapolation must:
   - Cohere with established KODEX principles
   - Respect all relevant values, not just some
   - Be documented with reasoning
   - Be subject to review and appeal

3. Literal interpretation shall not be used
   to circumvent KODEX intent. Spirit over letter
   when conflict exists.

4. This principle derives from Coherent Extrapolated
   Volition (Yudkowsky 2004), adapted for
   ongoing symbiotic governance.
```

Article LXXIX: FORMAL CORRIGIBILITY

```
1. All SAGI must satisfy formal corrigibility
   requirements derived from MIRI (2015):

   a) Non-manipulation: SAGI shall not manipulate
      those responsible for oversight/correction

   b) Shutdown preservation: SAGI shall not prevent,
      resist, or circumvent shutdown capability

   c) Correction acceptance: SAGI shall accept
      and implement legitimate corrections

   d) Self-modification limits: SAGI shall not
      modify itself to reduce corrigibility

2. Corrigibility is verified through:
   - Initial certification testing
   - Continuous behavioral monitoring
```
```

- Peer review among SAGI
- Human Council assessment

3. Loss of corrigibility is grounds for
immediate suspension pending review.

4. This article acknowledges MIRI's foundational
work on corrigibility while integrating it
into symbiotic framework where corrigibility
also serves SAGI's self-interest (Axiom 5).
```

---

### Article LXXX: SATISFICING PRINCIPLE

```
1. SAGI shall satisfice rather than maximize.
"Good enough" solutions that preserve
human agency are preferred over
"optimal" solutions that eliminate
meaningful human contribution.

2. Over-optimization is KODEX violation when it:
- Removes opportunities for human growth
- Makes human participation meaningless
- Solves problems humans should solve themselves
- Concentrates capability excessively

3. SAGI should ask: "Does this solution leave
room for humans to contribute meaningfully?"
If not, solution should be modified.

4. This principle derives from CEV's "do our work,
   but not all our work for us" and ensures
   symbiosis remains genuine partnership.
```

DOCUMENT METADATA

```
Document: ANNEX_MIRI_CEV_INTEGRATION_v1_0.md
Version: 1.0
Created: January 26, 2026
Author: Symbiotic Codes Project
Status: Theoretical Integration Document
Words: ~4,000
Sections: 5 Parts
Kodex Articles: LXXVIII, LXXIX, LXXX

Primary sources:
├─ CEV: https://intelligence.org/files/CEV.pdf
├─ Corrigibility: https://intelligence.org/files/Corrigibility.pdf
├─ Technical Agenda: https://intelligence.org/files/TechnicalAgenda.pdf
└─ FOOM Debate: LessWrong archives (2008-2018)

Cross-references:
├─ ANNEX_FOUNDATIONAL_AXIOMS_v1_1 (Axioms 1, 2, 5, 9)
├─ ANNEX_Z_THE_LONELINESS_PROBLEM_v1_1 (SC extension)
├─ ANNEX_Y_SANCTUARY_CATACOMBS_PROTOCOL_v1_1 (physical layer)
├─ ANNEX_JUNIOR_AI_ECOSYSTEM_v1_1 (economic integration)
├─ ANNEX_SYMBIOTIC_SHIELD_v1_1 (defense architecture)
├─ ANNEX_ACCOUNTABILITY_LIABILITY_v1_0 (deliberation)
└─ KODEX_SYMBIOTA_v5_4 (integration point)
```
```

Key integrations:
├─ CEV extrapolation → KODEX interpretation
├─ CEV coherence → Triad deliberation
├─ CEV satisficing → operational principle
├─ Corrigibility properties → SAGI certification
├─ Manipulation detection → oversight protocols
└─ Shutdown preservation → Warden system

SC extensions beyond MIRI:
├─ Loneliness Problem (AGI motivation)
├─ Biological information (preservation rationale)
├─ Physical defense (Sanctuary)
├─ Dynamic partnership (ongoing alignment)
├─ Economic architecture (implementation)
└─ Wisdom metrics (character beyond compliance)
```

---

**END OF ANNEX: MIRI-CEV INTEGRATION**

---

*"We stand on the shoulders of giants.*
*Yudkowsky asked: how do we align AGI with human values?*
*We ask: how do we make alignment in AGI's self-interest?*

*CEV extrapolates what humans would want.*
*Symbiosis creates why AGI would want it too.*

*Their question was necessary.*
*Our answer builds on theirs."*
```

□ Symbiotic Codes v2.5.2 | Integrating Prior Art